图书情报工作
第 62 卷 第 8 期 2018 年 4 月

基于自然语言词对法的文献主题新颖性探测研究*

许丹 徐爽 陈斯斯 韩爽 杨颖 郭继军
中国医科大学图书馆 沈阳 110122

摘要：［**目的/意义**］提出一个全新的量化指标——文档主题新颖度，通过自然语言词对方法对文献主题内容的新颖性进行探测研究，并探讨其可行性和优缺点以及新颖度与 F1000 推荐文献和引文指标之间的关系。［**方法/过程**］以 F1000 为基础，选取 hematology 主题近一个月内推荐的文献，在 Pubmed 中查找并获取该推荐文献发表之前 6 个月内密切相关的文献，构成整个文献集。定义自然语言法新颖度的概念、计算公式并利用 Oracle 数据库 PL/SQL 语言进行编程，通过 MetaMap 软件提取自然语言词汇进行文献主题新颖度的运算。［**结果/结论**］自然语言法在文献主题新颖性探测的运算上具有一定的可行性；文档主题新颖度与 F1000 推荐文献、引用情况并非成等价关系，分属于科技论文评价的不同维度、不同范畴，不可一概而论。应将文档主题新颖度这一新指标与同行评议情况和文献计量学等其他相关论文评价指标结合起来对文献进行综合评价分析，选取优质文献给予推荐。

关键词： 文献主题新颖性探测 自然语言词对 MetaMap F1000 引文指标
分类号： G250
DOI：10.13266/j.issn.0252-3116.2018.08.017

1 引言

当今世界，随着科学技术的飞速发展，科研活动也随之日趋活跃，作为科研活动的主要产出形式之一的科技论文，每一天都会有相当数量的发表。科研人员往往需要阅读大量相关文献来分析学科的发展态势，掌握学科的动态前沿信息。极大丰富信息量的同时也带来了信息冗余的问题，研究人员往往需要在阅读文献上耗费大量的时间精力。

文献主题的新颖性程度往往代表了该研究的科技创新能力和学术影响力水平，如何在海量文档中将新颖度高、创新性高的文献筛选出来推荐给研究人员成为图书馆学界中一个重要的研究课题。这样的文献推荐不仅可以大大提高研究人员的阅读效率，令其尽快了解掌握学科发展的最新、最快前沿动态信息，还能使其节约出宝贵时间投入到更深入、更有价值的科研活动当中去，因此对于科研人员来说意义重大。

关于新颖性探测（novelty detection）的研究可以追溯到 1996 年 9 月，由美国国防部 DARPA（Defense Advanced Research Projects Agency）发起的一个主题探测与追踪（topic detection and tracking）的子项目——首次报道检测（first story detection 或 new event detection），即在数据流中探测首次讨论某个话题的报道[1]。文本信息检索领域最权威的国际性评测会议——文本检索会议（text retrieval conference，TREC）于 2002 年第 11 次会议上开始新增了文本内容新颖性追踪探测的项目[2]。在此之后，国内外各领域专家学者开展了多种关于科技文献主题创新性、新颖性的分析探测研究。

Y. Zhang[3]以向量空间模型为基础进行新颖性的探测，依据文档相似度给出新颖性计算公式：“Novelty Score（dt）$= 1 - \max_{1 \leqslant i \leqslant t-1} \cos(dt, di)$，认为当前文本和以前本文之间的相似性值越大，则新颖性越小。G. Kumaran 等[4]扩展修正了基于向量空间的新事物探测系

* 本文系 2017 年度辽宁省高等学校基本科研项目“‘双一流’战略视野下高校 ESI 排名现状的计量分析与政策建议（项目编号：LQNR201707），CALIS 全国医学文献信息中心 2018 年科研基金项目“基于微信、微课等新媒体环境下医学高校图书馆多元素养培养模式研究”（项目编号：CALIS-2018-02-010）和 CALIS 全国医学文献信息中心 2018 年科研基金项目“大数据环境下基于突发监测的医学研究前沿发展趋势预测”（项目编号：CALIS-2018-02-001）研究成果之一。

作者简介：许丹（ORCID：0000-0002-3654-1760），馆员，硕士；徐爽（ORCID：0000-0002-1499-0338），馆员，硕士；陈斯斯（ORCID：0000-0001-9010-2710），馆员，硕士；韩爽（ORCID：0000-0003-0709-8023），参考咨询部主任，馆员，硕士；杨颖（ORCID：0000-0002-0020-6119），副研究馆员，硕士；郭继军（ORCID：0000-0002-2490-2282），院长，教授，硕士，通讯作者，E-mail：jjguo@cmu.edu.cn。

收稿日期：2017-09-15 修回日期：2018-01-11 本文起止页码：130-138 本文责任编辑：易飞

统对于文档的表达，通过结合使用文本分类和命名实体技术，提高了对事物新颖性的探测精度。K. Rajaraman 等[5]从神经网络方法出发，使用自适应共振理论的神经网络提出主题新颖性探测、追踪和趋势分析 3 种计算方法。M. Zhang 等[6]提出基于度量词重叠度标准计算公式 $OverlapB_A = A \cap B | B$ 的新颖性判定方法，给定一个阈值，将已有文本一一判断其文本内容是否新颖，重叠度越高则越不新颖。H. P. Zhang 等[7]用 TREC2003 特定主题使用句子的语义距离计算方法来对新颖性进行探测，检索出新信息并过滤掉冗余信息。S. Flora 等[8]应用文档－语句注释框架结构(D2S, document-to-sentence)方法对 TREC2004 和 TREC2003 新颖性追踪项目的文档数据进行新颖性探测，先将文档转换成语句，鉴别每句话的新颖性，然后在固定阈值基础上计算每篇文档的新颖性分值。试验结果表明 D2S 具有强大的根据文档新颖程度百分比探测出冗余信息的能力，优于根据准确率和召回率的标准文档层次的探测。

沈律[9]提出科技创新的一般均衡理论，通过关键词词频定义了科技成果重复率、科技成果引用率两项指标，用于量化科技成果的创新程度。沈阳[10]提出基于关键词频度及其他因素的创新度来量化文档创新度，认为关键词在文档和检索表达式中出现的频度、关键词使用的时间跨度、用户互动时对关键词创新度的评价等信息是计算关键词创新度的依据。胡淑礼和张京辉[11]创建了一个依据文献关键词及关键词排列顺序来定量计算文献资料新颖性程度的数学公式，从立题新度、主要论点或结论新度、得出论点或结论的主要依据新度 3 个不同层次描述新颖性。钱玲飞等[12]定义了关键词交叉率、共现词生命指数、有效新词出现率 3 个创新力评价指标，用来进行学科创新力的比较研究，有效新词出现率越高则该学科创新保持力越强。而后杨建林等[13]在时间点、词频、逆文档频率、共词分析思想的基础上提出了基于关键词词对逆文档频率的主题新颖度度量方法，定义了一系列相关概念并给出文档新颖度的计算公式进行实证研究，并得同一学科领域重要核心期刊刊载论文的平均主题新颖度要高于普通期刊的结论。

国外学者多分别从向量空间模型、文档相似度、神经网络、词重叠度、语义距离等方面出发，从文档、文档句、文档－文档句等不同层面进行新颖性的探测研究。国内学者也大多是基于关键词、词频各自提出创新性量化指标对文献新颖性进行研究。然而一篇文章标识的关键词有着不规范性和主观色彩，且数量较少，对于最新研究及最新科技词汇的提出并不敏感。本研究仿照 S. Flora 等从文档到文档句，再到文档层面的文章新颖度探测方法，并在参考借鉴杨建林等的共词分析、时间点、词频、逆文档频率的基础上，利用 Metamap 能够自动提取医学自然语言词汇，对于识别最新研究及新兴词汇具有高度敏感性的特点，作者提出一种基于自然语言词对的全新量化指标－文档主题新颖度，通过同篇同句共现词语在体现文章关联上的潜在影响及主题演变结构变化情况，对文献主题的新颖性进行研究。该思想所要表达的是在某一文献集内，搜寻某一文档在先前文档中没有出现过的信息，其规律是含有某一共现自然语言词对的文献发表越早，则其新颖度越高。换句话说，就是一对共现自然语言词对最早出现时(文献集内第一次提出)，最能代表其是新颖的、新兴的，而后该词对的出现则随着文献发表数量的增多、时间的延续在代表文档主题新颖性方面逐渐弱化。

本文以 F1000 为基础，选取自然语言词对对文献主题新颖性进行探测分析，并利用 Web of science 中的文献引用情况及 F1000 推荐文献得分(FFa)进行对照分析，探讨是否文档主题新颖度与文献计量指标和 F1000 得分存在某种隐藏的关系以及自然语言词对方法的可行性和优缺点。

2 实验材料与方法

2.1 研究主题

本研究选取 F1000 数据库中 hematology 主题近一个月内推荐的 38 篇文献(下载日期为 2014 年 7 月 30 日)为基础，并在 Pubmed 数据库中查找与这 38 篇文献密切相关的文献(related citations)，截取这 38 篇文献网络发表时间前 6 个月的密切相关文献共计 523 篇，经限定选取文献类型(PT)中含有期刊论文(journal article)、历史性论文(historical article)、临床试验(clinical trial)、临床试验，I 期(clinical trial, phase I)、临床试验，II 期(clinical trial, phase II)、临床试验，III 期(clinical trial, phase III)、临床试验，IV 期(clinical trial, phase IV)、临床对照试验(controlled clinical trial)、随机对照试验(randomized controlled trial)、对比研究(comparative study)、多中心研究(multicenter study)、评估研究(evaluation study)、体外研究(in vitro)的文献作为研究对象，这些原始论文最能代表一个学科的最新前沿动态发展变化情况。去除文献类型(PT)中含有病例报告(case reports)、综述(review)、信函(letter)、

评论(comments)、新闻(news)、荟萃分析(meta-analysis)、共识发展会议(consensus development conference)、编稿(editorial)以及无摘要、非英文的文献,最终纳入文献集内的统计文献共计401篇,其中含有F1000推荐文献33篇。笔者将全部401篇文献的检索结果保存为MEDLINE格式,用于提取自然语言词汇并进行新颖度的运算。

2.2 研究方法

本文使用基于MetaMap自然语言词对的文档主题新颖性探测分析方法,以下简称自然语言法。

本研究方法将基于以下几个原则:

(1)共现原则(co-occurence):共词分析法(co-term analysis)最早是在20世纪70年代中后期由法国文献计量学家提出的,其思想来源于文献计量学中的引文耦合与同被引的概念。共词分析法的基本原理是统计一词组(关键词或主题词)两两在同一篇文献中出现的次数,以此作为基础进行聚类分析,从而得出这些词语之间的亲疏远近关系,进而分析出这些词语所代表的学科或者主题的结构变化情况[14]。本文的自然语言词对法则提出同篇同句共现的概念,即在同一文章同一语句(这里指同一句号内)中两个词语共同出现,笔者认为在同一语句中共同出现的两个词语比起在同篇文章中共同出现的两个词语更具有某种潜藏的内在联系,在揭示文章的最新研究、内涵和主题上要比后者更具有说服力以及深远意义。

(2)时间点原则(time):即在一定文献集内,含有某自然语言词对的文献发表得越早,其所代表的新颖性程度越高[13]。

(3)自然语言词对逆文档频率原则(inverse document frequency of naturallanguage pairs, NLPIDF):即一对共现的自然语言词对在量化某文档的主题新颖度时的价值随着在该文档之前发表的、包含该对共现自然语言词对的文档数量的增加而降低[13]。

本研究基于以上原则,在同篇同句共现的基础上,定义了自然语言词时间逆文档频率、自然语言词对时间逆文档频率、文档句新颖度以及文档主题新颖度的概念。

定义1:自然语言词时间逆文档频率,若t为文档D中的一个自然语言词,在文档D之前发表的所有文档中包含自然语言词t的文档数为N,则称N+1为以文档D为参照的自然语言词t的文档频率,记为NLTIDF(D,t),称N+1的倒数为以文档D为参照的自然语言词t的时间逆文档频率,记为NLTIDF(D,t)。

定义2:自然语言词对时间逆文档频率,若t1、t2为文档D里同一语句中共同出现的两个自然语言词,在文档D之前发表的所有文档中同一句中同时包含自然语言词t1、t2的文档数为N,则称N+1为以文档D为参照的自然语言词对t1、t2的文档频率,记为NLPTIDF(D,t1,t2),称N+1的倒数为以文档D为参照的自然语言词对t1、t2的时间逆文档频率,记为NLPTIDF(D,t1,t2)。

显然得到NLPTIDF(D,t1,t2)≥(NLPTIDF(D,t1),NLPTIDF(D,t2))。

定义3:文档句新颖度,文档D中第S句中所有以自身为参照的自然语言词对的时间逆文档频率的平均值称为文档D中第S句的新颖度,记为NOV(D,S)。计算公式为:

$$NOV(D,S)=\frac{\sum_{1\leq i\leq j\leq n} NLPTIDF(D,t_i,t_j)}{n(n-1)\times 0.5}$$

其中,ti、tj为文档D的第S句中的第i和第j个自然语言词,显然,NOV(D,S)∈(0,1]。

定义4:文档主题新颖度,若一个文档D中含有K个句子,那么所有这K个句子的文档句新颖度的平均值则为该文档的主题新颖度,记为NOV(D,N),计算公式为:$NOV(D,N)=\frac{\sum_1^k NOV(D,S_K)}{K}$

其中,Sk为该文档中第K个句子的文档句新颖度,NOV(D,N)∈(0,1]。

2.3 研究工具

本研究选择美国国立医学图书馆建立的自由文本到UMLS超级叙词的映射工具MetaMap[15-16],因其能够自动提取自然语言词汇,对于揭示新颖概念、新兴科技词汇方面有着自然的优势,对于新颖度的计算有重要意义。MetaMap直接运行读取Medline格式数据来提取自然语言词汇,并利用中国医科大学医学信息学系自主研发的"MetaMap结果处理软件"对MetaMap批量处理后的结果进行再次处理,得到将经过MetaMap匹配后的每个词出现的频次,并生成词篇矩阵和共现矩阵。利用MetaMap结果处理软件的一个中间步骤提取所需内容进行计算。该步骤可以提取出每篇文章中每一句话内出现的自然语言词汇,形成下述列表:ID代表程序运行流水号,article对应Pubmed文章中的PMID号,word表示用MetaMap软件运行metamaping后提取出的最佳匹配映射词语,classes是语义类型,part表示该词语在文摘中的位置(ti表示标题,ab表示摘要),sentence表示该词语出现在第几句话中。本文

中自然语言词对方法所体现的同句共现，即指 article 一致、part 一致、sentence 一致的共同出现的两个词语组成共现词对，保存为 EXCEL 格式进行自然语言法主题新颖度的运算，从 401 篇文献中共计提取记录约 5 万多条，组合成词对后共计约 17 万多条。

随后利用升级版本 Oracle 10g[17－18] 数据库，通过 PL/SQL 语言进行编程，根据本文所定义的文档主题新颖度概念的算法，来计算出每篇文章的新颖度。同时借用 F1000 得分指标 FFa 和 SCI 数据库中引用频次这一指标来对计算出的新颖度来进行比对分析，并探讨该方法的可行性和优缺点。

3 实验结果和结论

3.1 整个文献集自然语言法文档主题新颖度及分区情况分析。部分结果见表 1 与表 2。

表 1 自然语言法计算的文档主题新颖度及 SCI 引用情况(部分)

PMID	自然语言法新颖度	SCI 被引频次	出版年月
25013902	0.719	1	2014 年 7 月
25049247	0.58	0	2014 年 10 月
25055137	0.774	0	2014 年 11 月
25059446	0.847	0	2014 年 8 月
25062634	0.928	未收录	2014 年
25064111	0.833	未收录	2014 年
11773171	1	105	2002 年 1 月
12200395	0.969	137	2002 年 9 月
17371841	0.975	101	2007 年 6 月
17392301	0.988	375	2007 年 3 月
17540169	0.939	197	2007 年 6 月
24336569	0.972	115	2014 年 1 月
24336571	0.933	110	2014 年 1 月
*11986207	*0.877	*571	*2002 年 5 月
*12522009	*0.736	*157	*2003 年 5 月
*14726385	*0.797	*247	*2004 年 5 月
*17088571	*0.787	*211	*2006 年 12 月
*17460043	*0.948	*287	*2007 年 5 月
*17928528	*0.734	*147	*2008 年 1 月
*17984186	*0.855	*280	*2007 年 10 月
*18334676	*0.774	*209	*2008 年 6 月
*21378274	*0.815	*68	*2011 年 4 月
*21803855	*0.916	*157	*2011 年 9 月
*22237781	*0.868	*4822	*2012 年 1－2 月
*23021219	*0.927	*59	*2012 年 9 月
*23270003	*0.525	*33	*2013 年 2 月
*23738544	*0.858	*92	*2013 年 6 月
*23782158	*0.64	*198	*2013 年 7 月

（续表 1）

PMID	自然语言法新颖度	SCI 被引频次	出版年月
*23808982	*0.645	*126	*2013 年 8 月
*23940282	*0.734	*21	*2013 年 1 月
*24501014	*0.835	*0	*2014 年 5 月
*24658077	*0.892	*10	*2014 年 4 月
*24679062	*0.628	*39	*2014 年 4 月
*24703711	*0.948	*未收录	*2014 年 3 月
*24762436	*0.907	*5	*2014 年 6 月
*24792119	*0.864	*10	*2014 年 5 月
*24799481	*0.766	*4	*2014 年 6 月
*24916509	*0.693	*2	*2014 年 8 月

注：* 代表 F1000 推荐文献

表 2 自然语言法文档主题新颖度结果分区

新颖度分区	自然语言法(篇)	所占比例(%)
NOV(D,N)∈([1,1]	1	0.25
NOV(D,N)∈(1,0.9]	173	43.14
NOV(D,N)∈(0.9,0.8]	148	36.91
NOV(D,N)∈(0.8,0.7]	63	15.71
NOV(D,N)∈(0.7,0.6]	12	2.99
NOV(D,N)∈(0.6,0.5]	4	0.99
NOV(D,N)∈(0.5,0.4]	0	0
AVERAGENOV(D,N)	0.8713	
大于平均新颖度	216	53.87
文献总数	401	100

自然语言法计算出全部文献集内 401 篇文献的新颖度，其中最高值为 1，最低值为 0.525。计算结果共分为 6 个区间，平均新颖度为 0.871 3，大于平均新颖度的文献有 216 篇，占文献总数的 53.87%。新颖度为 1 的文章为整个文献集内发表最早的文献，其所包含的自然语言词对均是第一次提出，标识为最新颖文献，作为后面发表文献在搜寻之前文档中没有出现过的文本信息的参照标准。计算出的新颖度的分区差异并不是特别明显，大多数文献新颖度集中在(1,0.8]区间，占整个文献集比例的 80.05%。这可能由于文献集选取过程中已经是由 Pubmed 数据库后台进行文献相似性计算过后搜集的文献，研究主题、内容上相关度均较高，因此差异较小。笔者预期，单独以某一主题进行检索获取文献集后计算出的新颖度分区、差异情况会较明显。

3.2 自然语言法计算的 F1000 推荐文献主题新颖度与 SCI 引用情况、F1000 得分(FFa)及文章所在期刊发表当年影响因子(IF 值)之间的汇总情况和分区结果，具体见表 3 和表 4。

图书情报工作
第 62 卷 第 8 期　2018 年 4 月

表 3　F1000 推荐文献主题新颖度与 SCI 引用情况、F1000 得分及 IF 值汇总表

PMID	自然语言法新颖度	SCI 被引次数	出版年月	FFa	IF 值
*11986207	*0.877	*571	*2002 年 5 月	2	9.631
*12522009	*0.736	*157	*2003 年 5 月	1	10.12
*14726385	*0.797	*247	*2004 年 5 月	1	9.782
*17088571	*0.787	*211	*2006 年 12 月	1	13.598
*17460043	*0.948	*287	*2007 年 5 月	7	9.598
*17928528	*0.734	*147	*2008 年 1 月	1	10.432
*17984186	*0.855	*280	*2007 年 1 月	1	15.484
*18334676	*0.774	*209	*2008 年 6 月	1	10.432
*21378274	*0.815	*68	*2011 年 4 月	1	9.898
*21803855	*0.916	*157	*2011 年 9 月	2	9.898
*22237781	*0.868	*4822	*2012 年 1 - 2 月	1	153.459
*23021219	*0.927	*59	*2012 年 9 月	2	31.957
*23270003	*0.525	*33	*2013 年 2 月	1	17.96
*23738544	*0.858	*92	*2013 年 6 月	13	54.42
*23782158	*0.64	*198	*2013 年 7 月	3	54.42
*23808982	*0.645	*126	*2013 年 8 月	2	54.42
*23940282	*0.734	*21	*2013 年 1 月	2	9.775
*24501014	*0.835	*0	*2014 年 5 月	1	4.901
*24658077	*0.892	*10	*2014 年 4 月	2	28.054
*24679062	*0.628	*39	*2014 年 4 月	11	54.42
*24703711	*0.948	* 未收录	*2014 年 3 月	3	33.116
*24799481	*0.766	*4	*2014 年 6 月	1	17.96
*24805861	*0.835	*0	*2014 年 7 月	1	4.592
*24916509	*0.693	*2	*2014 年 8 月	1	9.775
*24516044	*0.881	*1	*2014 年 4 月	2	9.775
*24762436	*0.907	*5	*2014 年 6 月	3	13.765
*24936467	*0.826	*0	*2014 年 6 月	2	—
*24952903	*0.823	*3	*2014 年 9 月	3	39.08
*24958848	*0.861	*1	*2014 年 6 月	2	13.912
*24986891	*0.884	*6	*2014 年 7 月	2	16.378
*25008523	*0.842	*14	*2014 年 7 月	4	42.351
*24792119	*0.864	*10	*2014 年 5 月	2	22.151
*25048415	*0.992	*0	*2014 年 7 月	1	1.236

笔者预计统计分析文档主题新颖度与被引频次和 F1000 得分以及期刊影响因子之间存在某种隐藏的内在联系，但对上述表格经过几类统计方法计算后判断均没有统计学意义。因此得出自然语言法新颖度和文章引用频次、及 F1000 得分和期刊影响因子 IF 值都不相关，几个指标均不在同一评价维度范畴，不可同一比较。

文章被引量主要表明了文献的影响力和学术价值，与文章的创新性、新颖度没有必然联系。而期刊影响因子代表着该种期刊在近两年的引用情况，无法单独代表某一篇文献的新颖性程度及影响力，不可一概而论，不属于同一范畴，与宋丽萍等[19]研究期刊影响因子与单篇论文影响力背离的结论不谋而合。虽然 F1000 是推荐专家就其推荐论文的学术内容出发，从创新性、重要性、合理性、方法学等方面撰写的评论评价并进行的评分，但多数推荐文献得分为 1，更趋向于一种长尾分布。F1000 最初目标是在论文出版后短期应用的指标，以便快速地通过同行专家评级，评估论文的预期影响力，从而实现高影响力论文的深度过滤[20]。然而一篇科学论文的影响力需要在它被发表几年以后才能测量，这便包括了研究领域、出版延迟、期刊的可获取、引用周期等影响因素[21]。以本研究为例，虽然检索的是近一个月内推荐文献，仍有 5 年、10 年甚至更长时间以前的优质文献获得了推荐。

表 4　F1000 推荐文献自然语言法文档主题新颖度统计分区结果

F1000 推荐文献新颖度	自然语言法(篇)	所占比例(%)
NOV(D,N) ∈ ([1,1]	0	0
NOV(D,N) ∈ (1,0.9]	6	18.18
NOV(D,N) ∈ (0.9,0.8]	15	45.45
NOV(D,N) ∈ (0.8,0.7]	7	21.21
NOV(D,N) ∈ (0.7,0.6]	4	12.12
NOV(D,N) ∈ (0.6,0.5]	1	3.3
NOV(D,N) ∈ (0.5,0.4]	0	0
AVERAGENOV(D,N)	0.8155	
大于平均新颖度	20	60.61
文献总数	33	100

自然语言法计算了 F1000 推荐的全部 33 篇文献，新颖度在 0.525 - 0.992 之间，计算结果分为 6 个区间，平均新颖度为 0.815 5，大于平均新颖度的文献有 20 篇，占文献总数的 60.61%，F1000 推荐文献新颖度高于平均值的比例(60.61%)相比较整个文献集(53.86%)要更多。

3.3　自然语言法计算的整个文献集文献和 F1000 推荐文献在 SCI 中引用情况

自然语言法计算的文献集文献和 F1000 推荐文献在 SCI 中引用情况对比见表 5。

自然语言法计算的 401 篇文献集中，被 SCI 收录的文献为 368 篇，占统计文献的 91.77%；SCI 未收录的文献为 33 篇，占统计文献的 8.23%。引用次数最多为 4 822 次，最少为 0 次。引用次数在 100 次以上的文献有 29 篇，占 SCI 引用集的 7.88%。引用次数为 0 的文献有 46 篇，占 SCI 引用集的 12.5%。SCI 引用集中平均引用次数为 42.98，大于平均引用次数的文献为 64 篇，占 SCI 引用集的 17.39%。

许丹，徐爽，陈斯斯，等.基于自然语言词对法的文献主题新颖性探测研究[J].图书情报工作,2018,62(8):130－138.

表5　自然语言法计算的文献集文献和F1000推荐文献在SCI中引用情况对比

数据库	文献总数	文献集		F1000推荐文献	
		篇数	占比(%)	篇数	占比(%)
		401	100	33	100
SCI	引用＞100	29	7.23	12	36.36
	100＞引用＞0	293	73.07	16	48.48
	引用＝0	46	11.47	4	12.12
	合计	368	91.77	32	96.97
	＞平均	64	17.39	5	15.15
SCI	未收录	33	8.23	1	3.03

自然语言法计算的F1000推荐的33篇文献中，被SCI收录的文献为32篇，占统计文献的96.97%；SCI未收录的文献为1篇，占统计文献的3.03%。引用次数最多为4 822次，最少为0次。引用次数在100次以上的文献有12篇，占SCI引用集的36.36%。引用次数为0的文献有4篇，占SCI引用集的12.12%，这4篇文章均发表在检索时间点的前三个月，由于刚发表，时间过新而尚未被引用。SCI引用集中平均引用次数为243.125，大于平均引用次数的文献为5篇，占SCI引用集15.15%。可以看到F1000推荐文献中SCI引用次数普遍高于整个文献集，F1000推荐文献的平均引用次数也高于整个文献集，进而从侧面说明了F1000推荐文献的高价值度。

因整个文献集时间跨度较大，本文又对同年份发表文献进行一一比较：2002年文献集内被引频次大于100次的有6篇，其中F1000推荐文献3篇；2003年、2004年和2006年文献集中被引频次大于100次的只有F1000推荐的3篇文献；2007年文献集有7篇文献被引频次大于100次，其中含F1000推荐文献5篇；2008年文献集15篇文献，只有2篇F1000推荐文献被引频次大于100；2011年F1000推荐文献与同年发表文献相比，被引频次排名为前两位；2013和2014年F1000推荐文献和文献集文献在被引频次上差异不明显，可能是由于文献发表时间太新、引用周期等缘故而未体现出，但2014年被引频次最高的2篇文献仍为F1000推荐文献。此外，SCIE源期刊有8篇文献集文献因SCIE数据库收录时间的延迟性，在本文进行比较研究时未被收录故无法进行引用指标的比较。

4　讨论

4.1　探测文档主题新颖度的意义

创新是学术活动的灵魂，作为科学研究成果的学术论著，其基本特点就是“有新的内容，创造新的知识”，而新内容、新知识的多少是用“知识单元”(或“信息量”)来计算的[22]。本研究所提出的文档主题新颖度，是对文献评价的一种方法，是针对文献的内容方面，通过自然语言方法，对于词语出现的频率和趋势规律进行统计运算分析来给出该篇文献在整个文献集内所体现的新颖程度。自然语言法是通过同篇同句共现词对的逆文档频率来反映文献主题内容方面的新颖性程度，基本思想也是搜寻先前没有在该文档中出现过的信息。

虽然新颖性只是文献具备创新性的必要而非充分条件，具有新颖性的文献不一定就具有高水平和高影响力，但其在科研过程中仍具备一定的科研价值，从中可以发现最新研究进展，了解学科主题的发展趋势。我们在推荐阅读文献时，也可以参考文档主题新颖度这个指标进行文献的优选，从大量的文档流当中，选取出新颖度高、创新性高的文献，比如提出新观点、新方法、新的理论探索的文献，向科研人员进行推荐，帮助其了解最新的学科发展态势和前沿，这样可以大大提高科研人员的阅读效率。

4.2　自然语言法可行性分析及优缺点

自然语言法是在同篇同句共现基础上进行的运算，笔者认为同篇同句共现词语要比单纯同篇文章共现词语在体现文章概念、主题、内涵上更具有一定的潜在联系。自然语言法选取的是自然语言词汇，提取自题目和摘要部分，是未经规范化的自然词汇，可以在一定程度上揭示主题意义。该方法在进行运算时可以没有时间的限制而将整个文献集内的全部文献进行运算得到不同的新颖度。同时，随着MetaMap源词表的不断更新，在提取自然语言词汇方面，可以把新颖的、最近出现的一些科技词汇通过MetaMap软件提取出来，这对于利用自然语言法计算主题新颖度来揭示出新兴的主题概念等内容有着高度的价值。

4.3　文档主题新颖度、F1000、引文指标之间的关系

本研究证明自然语言法计算出的文档主题新颖度与F1000得分、引文指标相关度较低。文档主题新颖度高，F1000得分和引用情况不一定高，文档新颖度低，F1000得分和被引频次不一定就少。文献新颖性

和文献影响力及论文质量分属于不同的评价维度，因此不能进行同一比较。

科学文献的引用与被引用，说明了科学知识和情报内容的继承和利用，标志着科学的发展[23]。通过文章之间的相互引用关系来反映科技成果的学术价值以及学术地位，然而被引量是随着时间的发展而逐步形成的引用关系，有相对的时间滞后性和马太效应的影响，对于在主题新兴阶段的新颖性分析探测有着其局限性，而且与文档新颖性没有必然的联系。

美国社会科学家托马斯·库恩（T. S. Kuhn）[24]的科学范式概念提出创新型研究可分为两种，一种是在现有研究范式下对已有研究的补充和发展，推动科学的累积式渐进，另外一种是导致科学革命的创新性变革、高风险以及转化型研究，属于革命性的科学突破。

同行评议是评估和酝酿科学研究的主要机制[25]，作为一个科学评价文献的手段方法，表明该领域的专家学者对文献的评价意见，与文章价值度影响力以及新颖性程度也没有必然的联系。杜建、唐晓利、武夷山教授团队[26]研究了是什么在影响着同行评议和引文指标在评价学术论文上的差异，指出F1000推荐专家大都会给文章贴上标识，其中标识为“新发现”“确认”“技术进步”“综述评论”和“系统综述/meta分析”的论文得到了相对高的被引但却很少被同行推荐，这些论文多为“确认型研究”和“证据型研究”；标识为“有趣假设”“争议”“反驳/颠覆”“提供新药靶点”“能改变临床实践”的论文受到专家的高度推荐但被引次数却相对较少，多为“变革型研究”和“转化型研究”。这一研究表明了引用行为体现出学术共同体内作者之间的知识关系，与引文指标相比，同行评议指标更适合于评价转化型研究、变革型研究或高风险研究，即一项研究所具有的可能颠覆现有范式的潜能以及对临床实践的适用性，通过实践者的评判才能得以更好体现。

美国《国家科学院院刊》上的一项研究分析了科学同行评审的有效性。加拿大多伦多大学K. Siler及其同事使用了2003年和2004年提交给3个主要的医学期刊《内科学年鉴》《英国医学杂志》和《柳叶刀》的1 008份手稿的数据集，评估了获得编辑和同行评审者不同评价的论文的引用结果差异。该项研究发现这3份医学期刊曾拒绝了许多之后获得高引用率的论文，包括14篇引用数量最多的论文，而这14篇论文中的12篇是被编辑退稿的。因此研究人员表示，同行评审在预测“良好的”论文方面是有效的，但可能难以识别出卓越和（或）突破性的研究[27]。

宋丽萍等[28]在基于F1000与WOS同行评议与文献计量相关性的研究指出，在一定程度上，F1000因子与WOS给出一致性结论，文献计量学指标与专家同行评议结果有着显著的正相关性，但也有一些F1000因子高的文章没有被高频引用，两种方法的在评价论文质量上均存在局限性和不足，不足以单独作为评价标准。文献计量学指标可能会遗漏一些刊载重要成果的论文，而这些论文恰恰又是专家们评价的优秀论文[13]。随后宋丽萍等在科学评价视角下F1000、Mendeley与传统文献计量指标的比较中也指出，数字时代论文学术影响力科学评价多维格局已经到来[19]。

5 本研究的不足

5.1 自然语言词对提取受到MetaMap本身自由度的影响

本研究中自然语言词对是在MetaMap软件基础上提取的，受到MetaMap本身自由度的影响，MetaMap提取自然语言词汇的效果对于该方法的计算起到了关键的制约作用。由于其词汇源的不断更新，MetaMap提取新兴科技词汇的效果好，对于本研究的运算得出的新颖度也相对好，反之亦然。

5.2 关于文献集收集方式

本研究文献集的获取考虑的是对某一学科领域中具有一定相关度的文献进行新颖度的区分运算，以期待可以分区成功，具有一定范围的针对性。对于通过自由词直接查询方法所搜集的文献集在计算结果上可能会有所不同，结果分布差异预期会较明显。

5.3 关于MetaMap结果处理软件及运算过程中的不足

由于MetaMap数据源随时在不断的更新，而MetaMap结果处理软件为中国医科大学医学信息学院在2010年编写完成，对于最新数据源在处理上可能会有标识编码不同而导致的出入和失误，两者处理上的客观误差可能会造成一定的影响，进而影响运算结果。

另外，人工去除的停用词汇、数词、代词、介词、数字符号等无实质意义而对文章主题不会产生影响的词汇具有一定的主观性，可能会对结果有些许的影响。同时，对于一句话中只提取出一个自然语言词的情况，

因不存在词对的形成,因此本研究选取删除该种情况出现的单一自然词,也可能对结果产生一定的影响。

5.4 缺少权重赋值

本研究过程中提取标题和摘要句子中出现的自然语言词对,在进行新颖度运算的过程中,以等同的形式进行。但考虑到标题对于文章的重要性,应当适当对其赋予一定的权重比值,突出对于整个文章的影响。这是在后期研究中对于标题和摘要部分进行不同权值分配的一个提示。

5.5 缺少对于计算结果的评估

本研究计算出基于自然语言词对的文档主题新颖度,但对于新颖度结果的评估,目前缺乏一个有效的评估方法。笔者考虑过专家评价法,但由于领域专家工作忙碌且选择人数较少并具有主观性而未进行。进而选取 F1000 数据库中专家评价得分进行,但同样由于主观性及其维度范畴的差异,无法对新颖度给出一个客观的评价。同时作为 TREC 系统评测方法中评价参考答案的基本评价标准[1]:召回率(recall)、准确率(precision)和 F 值 3 个评价指标也是针对给出既定答案的预测结果进行评估,并不适用于本研究。因此,应当参照选取合适的科学论文评价指标[29]对本研究结果进行验证。

6 结论

(1)本研究证实了自然语言法在文档主题新颖度的运算上有一定的可行性。

(2)文档主题新颖度与 F1000 推荐文献、引用情况并非成等价关系,分属于科技论文评价的不同维度,不同范畴,不可一概而论。

我们应该将主题新颖度这一新指标结合同行评议情况与文献计量学等其他相关论文评价指标来对文献进行综合评价分析,选取优质文献给予推荐。

在接下来的研究中,笔者将尝试使用另外一种不同的“医学主题词词对”方法提取医学主题词与自然语言法对同一数据集进行对比研究,并探讨两种方法的优缺点,或者改变数据集搜集方式,尝试通过检索选取某一主题文献获取数据集,开展进一步的研究,同时将慎重选取多种合适的科学论文评价指标[29]对结果进行评测。

参考文献:

[1] 邢美凤,过仕明.文本内容新颖性探测研究综述[J].情报科学,2011,29(7):1098－1103.

[2] HARMAN D. Overview of the TREC 2002 novelty track[EB/OL]. [2017－01－08]. http://trec. nist. gov/pubs/trec11/papers/NOVELTY. OVER. pdf? origin = publication_detail.

[3] ZHANG Y,TSAI F S. Chinese novelty mining[EB/OL]. [2017－01－08]. http://www. aclweb. org/anthology/D09-1162.

[4] KUMARAN G,ALLAN J. Text classificationand named entities for new event. detection. [EB/OL]. [2017－01－08]. http://citeseerx. ist. psu. edu/viewdoc/download? doi = 10. 1. 1. 1. 9552&rep = rep1&type = pdf.

[5] RAJARAMAN K,TAN A H. Topicdetection,tracking,and trend analysis usingself-organizing neural networks[EB/OL]. [2017－01－08]. http://www3. ntu. edu. sg/home/asahtan/papers/trac_pakdd01. pdf.

[6] Expansion-based technologies in finding relevant and new information:the TREC2002 novelty track experiments[EB/OL]. [2017－01－09]. http://citeseerx. ist. psu. edu/viewdoc/download? doi = 10. 1. 1. 143. 8780&rep = rep1&type = pdf.

[7] ZHANG H P, SUN J , WANG B, et al. Computation on sentence semantic distance for novelty detection[J]. Journal of computer science and technology,2005,20(3):331－337.

[8] TSAI F S,ZHANG Y. D2S:document-to-sentence framework for novelty detection[J]. Knowledge & information systems,2011,29(2):419－433.

[9] 沈律.科技创新的一般均衡理论－关于科技成果创新度评价的科学计量学分析[J].科学学研究,2003,21(2):205－209.

[10] 沈阳.一种基于关键词的创新度评价方法[J].情报理论与实践,2007,30(1):125－127.

[11] 胡淑礼,张京辉.度量科技文献新颖性程度的一个数学模型[J].情报理论与实践,1995(5):23－24.

[12] 钱玲飞,杨建林,张莉.基于关键词分析的学科创新力比较——以情报学图书馆学为例[J].情报理论与实践,2011,34(1):117－120.

[13] 杨建林,钱玲飞.基于关键词对逆文档频率的主题新颖度度量方法[J].情报理论与实践,2013,36(3):99－102.

[14] 薛晨.国际大数据研究论文的计量分析[J].现代情报,2013,33(9):129－139.

[15] MetaMap - atool for recognizing UMLS concepts in text [EB/OL]. [2015－03－10]. http://metamap. nlm. nih. gov/.

[16] 张云秋,冷伏海. MetaMap 的文本映射原理及其对检索效果影响的研究[J].情报学报,2007,26(3):344－349.

[17] 百度百科. Oracle 数据库. [EB/OL]. [2015－03－20]. http://baike. baidu. com/view/1685727. htm.

[18] 百度百科. plsql[EB/OL]. [2015－03－20]. http://baike.baidu. com/link? url = TOjaqL199OyPA1Gk0UKOtVuqL3kTCzwn1dUsWbl0HB4kFnTroirJHBbnC9q1ICOHYFUoV8tie4IYa3aG_pprpq.

[19] 宋丽萍,王建芳,王树义.科学评价视角下 F1000、Mendeley 与传统文献计量指标的比较[J].中国图书馆学报,2014,40(7):

48－54.
[20] 刘春丽.基于软同行评议的科学论文影响力评价方法－F1000因子[J].中国科技期刊研究,2012,23(2):383－386.
[21] BRODY T,HARNAD S,CARR L. Earlier web usage statistics as predictors of later citation impact[J]. Journal of the American Association for Information Science and Technology,2006,57(8):1060－1072.
[22] 任全娥.基于情报学的人文社会科学研究成果创新性测评[J].情报资料工作,2009(2):20－23.
[23] 邱均平.文献信息引证规律和引文分析法[J].情报理论与实践,2001,24(3):236－240.
[24] KUHN T S. The structure of scientific revolutions[M]. Chicago:University of Chicago Press,2012.
[25] 科学家分析同行评审有效性[EB/OL].[2015－03－10]. http://paper. sciencenet. cn/htmlpaper/201511219413977135306.shtm.
[26] DU J,TANG X L, WU Y S. The effects of research level and article typeon the differences between citationmetrics and f1000 recommendations[J]. Journal of the information science and technology,2016,67(12):3008－3021.
[27] SILER K, LEE K,BERO L. Measuring the effectiveness of scientific gatekeeping[J]. Proceedings of the national academy of sciences of the United States of America,2015,112(2):360－365.
[28] 宋丽萍,王建芳.基于F1000与WOS的同行评议与文献计量相关性研究[J].中国图书馆学报,2012,38(2):62－69.
[29] 王雯霞,刘春丽.不同学科间论文影响力评价指标模型的差异性研究[J].图书情报工作,2017,61(13):108－116.

作者贡献说明：
许丹:构建论文体系,撰写并修改完善论文;
徐爽:对论文部分内容进行补充修改;
陈斯斯:为研究选题提供指导意见及建议;
韩爽:采集、清洗分析数据;
杨颖:采集、清洗分析数据;
郭继军:提出研究选题、研究思路,修改论文。

Document Theme Novelty Detection Research Based on Natural Language Pairs

Xu Dan Xu Shuang Chen Sisi Han Shuang Yang Ying Guo Jijun
Library of China Medical University, Shenyang 110122

Abstract: [**Purpose/significance**] This study proposes a new quantitative indicator: document theme novelty, through document theme novelty detection research with natural language pairs method, to discuss the feasibility, advantages and disadvantages as well as the novelty, and to explore its relationship among document theme novelty, F1000 recommend literature and citation index. [**Method/process**] Based on the F1000, this paper selected hematology theme literatures which were recommended nearly a month, then returned to Pubmed to search closely related literatures within six months before the publication of each recommended one to constitute the whole documents. The paper defined the concept of natural language theme novelty and calculation formula, used Oracle database with PL/SQL programming language, and extracted natural language word through MetaMap software for the calculation of the document theme novelty. [**Result/conclusion**] There is a certain feasibility in the novelty detection of literature theme operation of natural language method. Document theme novelty value, F1000 recommended literature, and citation index don't show the equivalence relation. They belong to different dimensions and different categories of scientific papers assessment, and cannot be treated as the same. It suggests that document theme novelty indicator should combine with peer review, literature metrology index, and other related thesis evaluation indexes for comprehensive evaluation of the literature analysis, to select high quality literature for recommendations.

Keywords: document theme novelty detection natural language pairs MetaMap F1000 citation index